2020

A Deep Learning Approach to Mapping Irrigation: U-Net IrrMapper

Thomas Henry Colligan IV
University of Montana

A DEEP LEARNING APPROACH TO MAPPING IRRIGATION: U-NET IRRMAPPER

By

Thomas Henry Colligan IV

Bachelor of Arts, The University of Montana, Missoula, MT, 2018

Thesis

presented in partial fulfillment of the requirements
for the degree of

Master of Science
in Computer Science

The University of Montana
Missoula, MT

December 2020

Approved by:

Ashby Kinch Dean of The Graduate School
Graduate School

Douglas Brinkerhoff Ph.D., Chair
Computer Science

Jesse Johnson Ph.D.
Computer Science

Marco Maneta Ph.D.
Geosciences

ABSTRACT

Colligan, Thomas Henry, M.S., December 2020 Computer Science

A Deep Learning Approach to Mapping Irrigation: U-Net IrrMapper

Chairperson: Douglas Brinkerhoff

Accurate maps of irrigation are essential for understanding and managing water resources in light of a warming climate. We present a new method for mapping irrigation and apply it to the state of Montana over the years 2000-2019. The method is based on an ensemble of convolutional neural networks that only rely on raw Landsat surface reflectance data. The ensemble of networks method learns to mask clouds and ignore Landsat 7 scan-line failures without supervision, reducing the need for preprocessing data or feature engineering. Unlike other approaches to mapping irrigation, the method doesn't use other mapping products like the Cropland Data Layer or the National Land Cover Dataset, removing the biases inherent in using those products. We evaluate our method and compare it to existing maps of irrigation on novel spatially explicit ground truth data, finding that our method outperforms other methods of mapping irrigation in Montana in terms of overall accuracy and precision. We find that our method agrees better statewide with the USDA National Agricultural Statistics Survey estimates of irrigated area compared to other methods, and has far fewer errors of commission in rainfed agriculture areas. In addition, our method produces uncertainties for predictions of irrigated land, and we find that the neural networks have large uncertainty in some misclassified areas. The methodology has the potential to be applied across the entire United States and for the complete Landsat record.

Introduction

Irrigated cropland is responsible for approximately 40% of the world's crop production, despite only comprising 20% of all cropland by area (FAO 2020). Irrigation already makes up about 85-90% of all consumptive freshwater use globally (Deines, Kendall, Crowley, et al. 2019), and agriculture accounts for around 80% of consumptive use in the United States (USDA ERS, 2019), of which use for irrigation is a large percentage. The FAO forecasts that by 2030 the world will have increased irrigated land by 34% (FAO 2020). Despite the positive impact irrigation has on food production, it's associated with a range of environmental impacts ranging from aquifer and surface water depletion to soil salinization. The reliance of agriculture on irrigation makes spatially explicit knowledge of when irrigation is occurring and how it's changing over time essential to understanding and managing water resources. Maps of irrigation can also help inform climate and earth systems models, for example providing estimates of water diverted for agriculture to groundwater models. Most large scale maps of irrigation have been produced at low resolutions and/or rely on remote sensing data combined with national or subnational statistics (Thenkabail et al. 2009, Siebert et al. 2015, Meier, Zabel, and Mauser 2018, Salmon et al. 2015, Brown and Pervez 2014, Xie et al. 2019, Ozdogan and Gutman 2008). Statistics collected by governments or NGOs often do not capture high resolution temporal trends in irrigation due to the labor-intensive collection of data and can contain self-reporting bias, especially if there are political or other incentives involved in reporting irrigation. Methods based solely on census data are limited in resolution by the administrative boundaries used to collect the data. For example, the United States' National Agricultural Statistics Survey (NASS), which reports irrigated area by county in the US, is compiled every 5 years by self-reported survery. The statistics are generally regarded to be an accurate characterization of crop type and irrigation status for lands across the US, but could be biased due to incentives to over- or under-report (Ozdogan and Gutman 2008; Deines, Kendall, Crowley, et al. 2019). This occasional snapshot of irrigated area does not provide the temporal resolution required to pick up short term fluctuations in irrigated agriculture. In addition, NASS excludes irrigated areas that produce less than $1,000 (USDA 2020), slightly biasing its surveys. Definitions of irrigation may also depend on the agency collecting the data, and may provide area equipped for irrigation, not actual irrigated area (Ozdogan and Gutman 2008). Irrigated area and area equipped for irrigation may be significantly different in any given year due to fluctuations in crop prices or rotation. The spatial resolution (ranging from 500m to 5 arcsec) of many existing maps of irrigation of the globe and the continental United States (CONUS) also makes the identification of small or partially irrigated areas difficult. The low resolution of previous maps of irrigation on large scales is due to the extensive computing resources required to process satellite data and the effort required to gather ground truth samples. However, the increasing ability of researchers to work with large data sets has made mapping irrigation at high resolution possible.

Platforms like Google Earth Engine (GEE) (Gorelick et al. 2017) have aggregated remote sensing data in one place and provided free access to their compute infrastructure, helping researchers increase the resolution and scale of remote sensing mapping approaches. The introduction of GEE to the remote sensing community has made mapping irrigation at high resolution on large temporal and spatial scales possible without relying on local computing power. This shift in technology has led to multiple efforts at mapping irrigation using Landsat data on large scales and high resolutions (Deines, Kendall, and Hyndman 2017; Deines, Kendall, Crowley, et al. 2019; Xie et al. 2019; Ketchum et al. 2020). These efforts all rely on Random Forest (RF) models and novel ground truth datasets to map irrigation across regions ranging from the High Plains Aquifer to the entire continental United States (CONUS). RFs are common in the remote sensing community for their flexibility, interpretability, and ease of use. They also are able to handle high-dimensional and complicated data. These are the only high resolution (30m) irrigation maps that cover large regions in time and/or space in the United States. LANID-2012 (Xie et al. 2019), is a Landsat-based irrigation map that uses a semi-automatic method of generating training data to train RF models that classify irrigation across CONUS at 30m. The semi-automatic generation of training data uses estimates of irrigated area by county from MIrAD-US (Brown and Pervez 2014) (a 250m irrigation map that matches NDVI- derived estimates of irrigation area directly with NASS statistics by county) to threshold maps of satellite-derived spectral indices until the area reported by the maps and MIrAD match. The thresholded maps serve as areas in which (Xie et al. 2019) sample points to train RF models. The trained RF models are used to classify land as irrigated or non-irrigated. LANID-2012 also incorporates some hand-labeled training points in humid areas over the US to ameliorate the failures of their automatic extraction of training data in wet areas. Their

method is still based on county level statistics through the reliance on training data generated by matching MIrAD-US. IrrMapper (Ketchum et al. 2020) and AIM-HPA (Deines, Kendall, Crowley, et al. 2019), take a different approach and train their random forest models only with hand-labeled training points, producing estimates of irrigated land independent of NASS. IrrMapper covers the western United States and AIM-HPA extends across the High Plains Aquifer. LANID and AIM-HPA both mask their classification product to an ancillary land cover map such as the Cropland Data Layer (CDL). IrrMapper (and our method) removes this postprocessing step, avoiding the biases included with the application of another map product. The RF methods above (LANID, IrrMapper, and AIM-HPA) all reports overall accuracy greater than 90% and agree well with NASS statistics by county.

Part of RFs widespread acceptance is their free implementation in GEE. However, the state of the art in image segmentation (of which irrigation mapping is a subset) is now achieved through the application of deep convolutional neural networks (CNNs), not through RFs. CNNs can match and even exceed human performance in various computer vision tasks and are the principal approach to image classification and segmentation. CNNs incorporate learned representations of images on many scales when making predictions, picking up patterns and contextual information. This is especially important in irrigation mapping, where the difference between a non-irrigated field and irrigated field may not be solely in spectral information but in the shape of the field or surrounding context. RF models, in contrast, have to rely on hand-derived features that capture spatial context, or otherwise discard spatial information in their classification. Another advantage of CNNs is that they can use dense pixel-wise labels while RF models have to pick and choose which points they use for training. CNNs have been used in many domains of remote sensing, from crop type mapping and land cover classification (Zhong, Hu, and Zhou 2019; Rußwurm and K rner 2018), to change detection (Peng, Y. Zhang, and Guan 2019; Janalipour and Taleai 2017). Despite their proven utility in remote sensing applications, there are only a few examples of CNNs directly applied to producing maps of irrigation, especially at large scales. The main reason for this deficit is the limited availability of training data (Deines, Kendall, and Hyndman 2017) and high computational costs. Training deep neural networks requires thousands of training, testing and validation samples as well as specialized computer hardware (e.g. graphical processing units, GPUs). Evaluating these models over a large geographic region at high resolution is also cost-prohibitive, as each application of a CNN requires millions of floating-point operations and is not readily parallelizable. Despite the disadvantages, the accuracy of CNNs routinely surpasses that of RF-based approaches because of their ability to learn explanatory features without supervision. The only use of deep learning to produce maps of irrigation to our knowledge at the time of writing is the work by (C. Zhang et al. 2018), and (Saraiva et al. 2020). Both methods use CNNs to map center pivot irrigation patterns over the state of Colorado and the Cerrado Biome of Brazil, respectively. These methods both relied on intensive hand-labeling of center pivot irrigation structures in satellite images. Both achieved overall accuracy of greater than 99%, and limited their study area to a relatively small region in time and space (20000km^2, 1986 and 2000, and 37000 km^2, 2017, 2018 and 2019, respectively).

Our objective was to produce a neural network model and framework that is easy to use and achieves better results than current approaches to mapping irrigation on a large spatial and temporal scale. We use novel irrigated ground truth data (a subset of the data used in Ketchum et al. 2020) to train an ensemble of U-Net architectures (Ronneberger, Fischer, and Brox 2015) to label Landsat pixels as irrigated or non-irrigated in the state of Montana for years 2000-2019. In addition, we use the model ensemble to produce pixel-wise uncertainties that agree with qualitative assessments of uncertainty. Montana was chosen as the study area because it covers a large geographic area (380,000km^2) and has many different types of irrigation. We conduct a pixel-wise comparison between methods (IrrMapper, U-Net, and LANID) for mapping irrigation using the novel ground-truth data, a spatially explicit evaluation not performed for other products. Our method outperforms these other methods in overall accuracy and errors of commission and omission, despite limited feature engineering and the omission of a crop mask like the CDL. Finally, we compare our method to NASS statistics and Final Land Unit (FLU, Revenue 2015) classification data from Montana. FLU is a vector dataset that classifies private agricultural land into one of six uses: fallow, hay, grazing, irrigated, continuously cropped, and forest. The data are used for property valuation for agriculture on private properties. We find better agreement for most years between total irrigated area in Montana calculated with our approach and NASS data compared to existing maps of irrigation. We also report on the qualitative performance of U-Net, LANID, and IrrMapper, describing the advantages and drawbacks of each product.

Methods

Model description

We used a variant of the U-Net model (Ronneberger, Fischer, and Brox 2015) to perform semantic segmentation of irrigated areas. Fully-convolutional neural networks (FCNNs) such as U-Net are noted for their fast inference time compared to pixel-wise methods (Shelhamer, Long, and Darrell 2017), which means maps of irrigation can be produced in little time on a personal computer once the network is trained. The general U-Net architecture was chosen because it has been established as a tried-and-true method for image segmentation. A diagram of the model is shown in Fig. 1, and coincidentally is very similar to the architecture used in (Saraiva et al. 2020) to identify center pivot irrigation. The model was composed of two main parts: A contracting path that is able to extract low-level features and spatial context, and an expanding path that incorporates low-level information from the contracting path to produce high resolution predictions. We followed each convolutional layer with a batch normalization layer (Ioffe and Christian Szegedy 2015) and used the ReLU activation function. We used weight decay with a coefficient of 0.001 on all convolutional layers. Unlike the original U-Net implementation, we used 2-d upsampling instead of transposed convolution to increase resolution. Each convolution is zero-padded. The final layer in the model is a convolutional layer with kernel size 1 and depth 3 followed by a softmax activation function. The softmax layer produces three predictions of class probability per pixel that sum to 1. In the original implementation of U-Net, the number of filters at each convolutional layer starts at 64 and doubles after each max-pooling layer. We use the same approach, but the first convolution in our U-Net implementation consists of only 32 filters, reducing the overall number of parameters from 32 million to 8 million. Fewer parameters means the model took less time to train and evaluate.

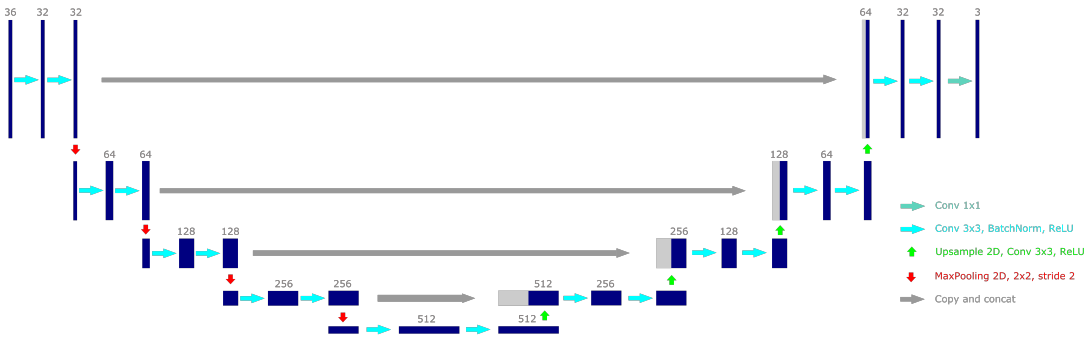

Figure 1: Neural network architecture used in this study. Blue bars indicate convolutional filter blocks. Numbers above blocks represent the number of filters in the block. Arrows represent operations on the blocks.

Many applications of U-Net perform segmentation on single images, with either 3- or 4-channels. Using a single image for land cover classification would discard important temporal information, which is especially important for classifying irrigation (Ozdogan and Gutman 2008). Popular deep learning approaches aimed at classifying time-dependent satellite data include recurrent neural networks and attention based approaches (Rußwurm and K rner 2018, Garnot and Landrieu 2020, Rußwurm and Körner 2019; Xu et al. 2020). RNNs require a forward pass through the model for each token of the input sequence. In land cover mapping applications, a "token" is a single image, or stack of images captured at a single date. This means each forward pass through the model is expensive. Self-attention operations are also notoriously resource hungry, with computational costs scaling quadratically with the number of elements in the input sequence. The costs associated with an explicitly temporal model were deemed too high for large scale spatiotemporal evaluation. The idea of concatenating features from different times and feeding them to a non-temporal model is not new (Ozdogan, Yang, et al. 2010) yet has not been applied in a fully-convolutional neural network. Accordingly, we concatenate images taken over a growing season into a 36 channel image. The 36 channel image consists of 6 images with 6 channels collected at different times.

Training details

We trained 10 models on random subsets of the training data as model ensembles are shown to do better than their constituents and can provide a measure of predictive uncertainty (Breiman 1996). Each model was trained for 100,000 gradient descent steps (batch size 25) with the Adam (Diederik and Ba 2014) optimizer with default α, $\beta_{1,2}$ parameters. We used random majority undersampling to roughly balance examples from each class per batch. The initial learning rate was 0.001, and was step decayed every 25,000 steps by 60%. Each model took around 24 hours to train on a single NVIDIA GTX 1070 GPU.

Study area

The study area is the state of Montana (MT), United States. West of the Continental Divide (CD), Montana is generally wetter and more temperate than area east of the CD, which is drier and windier (DNRC 2015). Montana's mean annual precipitation for 2000-2019 is shown in subfigure b) of Fig. 2. Irrigation in the state of Montana is mostly applied by diverting surface water (99%) and groundwater pumping (1%) (Lonsdale et al., 2020), and comprises ∼68% of consumptive water use (DNRC 2015).

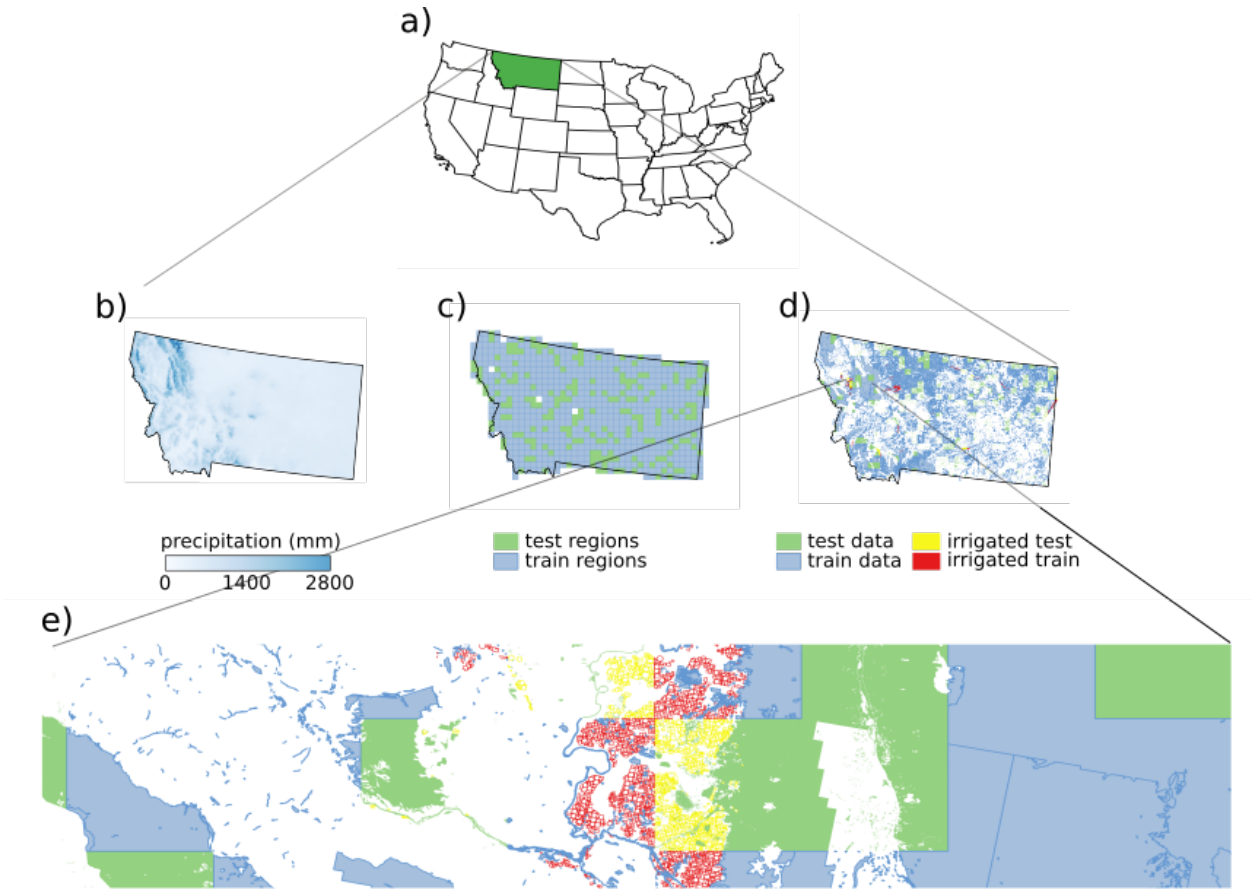

Figure 2: Study area and data used in this analysis. Subfigure b) indicates the mean annual precipitation for the years 2000-2019, c) shows the test and train regions spread over the state of Montana, d) the training and test polygons. A zoomed in view of the training and test polygons is shown in subfigure e. Irrigated training polygons are in red and test polygons are in yellow. For ease of visualization, the unirrigated and uncultivated class are aggregated into one class.

Labels

We used the data in (Ketchum et al. 2020) restricted to the state of Montana. Irrigated field labels in the data were produced through manual interpretation of satellite images, along with ancillary information. The training data consisted of three classes: irrigated, unirrigated, and uncultivated. Unirrigated data consists of predominately dryland agriculture with some fallowed fields. The uncultivated class is land which are not used for crop cultivation, and represents rivers, lakes, wetlands, forest, and prairie. Irrigated and fallow field labels were collected for 2003, 2008-2013, and 2015 to represent variability in water availability, using yearly precipitation as a proxy. Areas labeled as unirrigated or uncultivated were assumed to stay the same class over the years studied. Overall, training data was used to train the model for years 2003, 2008-2013, and 2015. The vector field labels were rasterized and downloaded from GEE along with corresponding Landsat data to train the models.

To ensure no correlation between the training and validation data, a grid of 776 23040 m^2 non-overlapping squares was spread over Montana and then split into subsets of 582 and 194 (80, 20% splits, subfigure c in Fig. 2). The vector labels described above were then split into training and validation sets based on this grid. A test set was not used in this analysis because hyperparameters were not tuned on the validation set, as meta-optimization is often costly. We wanted to create a framework that didn't need exhaustive tuning. Training features were sampled at random points within the training labels, resulting in repeated training samples. The validation set did not contain any repeated samples.

Features

We used Landsat Tier 1 Surface Reflectance data from Landsat missions 5, 7 and 8, acquired through GEE. The Landsat mission is the longest continuous record of satellite images of the earth, and minimal effort is required to standardize across missions. Landsat images of the earth are also captured relatively frequently, with approximately 8 days between subsequent observations. The bands used for analysis were blue, green, red, NIR, SWIR 1, and SWIR 2. We did not perform any cross-calibration between Landsat sensors. Some land cover mapping tasks remove clouds, snow, and cloud shadow from their classification product, and some interpolate missing data due to clouds based on surrounding noise-free pixels. Cloud masks (and corresponding interpolation of missing data) were not used in the training data in order to force U-Net to learn a robust representation of irrigation that ignores clouds.

To standardize data across years, we took temporal means of images captured over a 32-day period. We assumed that the temporal mean value of a Landsat band will stay more stationary over time than the values of the individual bands at a certain capture date. Temporal aggregation of features also reduced the amount of data necessary to train the model. All available Landsat images starting on May 1st and extending for 192 days were collected and sorted by time using GEE. For each non-overlapping 32-day segment within the 192 days, the temporal mean of all images collected during that segment was computed, and the resulting mean image was stored and used as a feature to the model. The final features that were ingested into the model were six "temporal mean" images sorted by time.

We treat each band as a separate feature without temporal information, which is standard for image segmentation models. The 32-day time period was chosen because it represented a trade-off between temporal resolution and number of input features. Taking the mean over a longer time period could suppress information important for determining irrigation. A shorter aggregation period would cause the number of input features to rise, increasing computational and storage needs. Another problem in the training and validation data is the presence of the Landsat-7 scan-line failure. The failure occurred on May 31, 2003, so the scan-line failure was present in all of the training and validation data. Temporal averaging smoothed the missing scan lines somewhat, but the missing data was noticeable in the data upon inspection. Some scan line artifacts are visible in the final product, but on the whole U-Net learns to ignore clouds and scan line failures. The noisy data serves as inherent data augmentation, as U-Net has to learn a representation of irrigation that ignores clouds and scan line artifacts.

Analysis Pipeline

Maps of irrigation were produced for the years 2000-2019 over Montana. We used the overlap-tile strategy in (Ronneberger, Fischer, and Brox 2015) to produce seamless predictions over the region of interest for each

model. Each model predicted a probability for each class (irrigated, uncultivated, unirrigated) for each pixel, where the maximum probability indicates which class the model predicts the pixel to be. The probabilities produced by each model are floating point numbers between 0 and 1. To reduce data storage requirements, the output of the models was scaled to the range 0-255 and converted to an unsigned 8-bit integer type. The loss of resolution from this conversion was negligible. The median predicted probability for each pixel in each band was calculated and the largest median value in each class was used as the final classification. To quantify ensemble uncertainty, we created maps of inter-quartile range (IQR) of the predictions for each pixel. A smaller IQR meant that models agreed well on the classification of a given pixel, and a large IQR indicated high disagreement. The IQR maps were calculated after the conversion of the softmax probabilities to 8 bit integers.

The vector labels used to train the model were collected at higher resolution (centimeter to meter resolution imagery) than the Landsat resolution, resulting in separate fields that were close to one another becoming one shape when rasterized. Accordingly, the models learned to label small roads between fields as irrigated. An exterior buffer was applied to the vector labels to resolve this problem but was unsuccessful. An approach that heavily penalizes misclassifications near borders of labeled objects (similar to the one in (Ronneberger, Fischer, and Brox 2015) was also considered. The borders of our labeled objects aren't

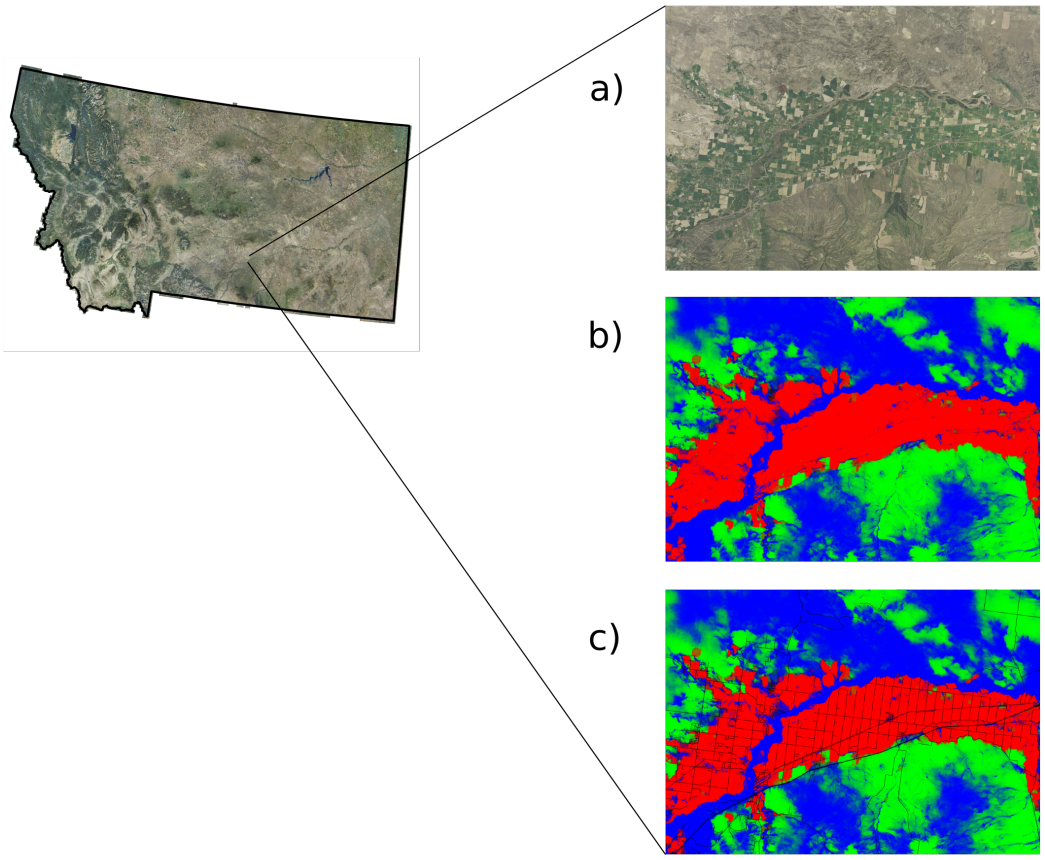

Figure 3: a) NAIP imagery of irrigation near Billings, Montana (2013). b) corresponding unmasked model predictions. Red indicates irrigation. c) Irrigation predictions masked to roads shapefile.

assigned a label, unlike in the original U-Net implementation, where borders were assigned to the background class. The border class in our problem could represent a wide variety of land cover classes including irrigation, and incorporation may decrease model accuracy. Instead of trying to force the models to learn a "road" or "border" class, we used human-labeled data to post-process roads out of model predictions. We masked the pixels in the final evaluated rasters with vector features that describe all roads in Montana

(downloaded from https://geoinfo.msl.mt.gov/home/msdi/transportation/). An example of the final map is shown in Fig. 3.

Accuracy analysis

We conducted a pixel-wise comparison of LANID, IrrMapper, and U-Net on the validation data. The validation data consisted of hundreds of millions of pixels, with ∼100× more non-irrigated pixels than irrigated pixels. We generated results for the three classes predicted by U-Net for all years with irrigated labels (2003, 2008-2013, and 2015). The comparison between the three available maps was only on their performance on data consisting of only irrigated and non-irrigated classes. F1-score was used as a metric to quantify errors of omission and commission on irrigated labels.

We compared U-Net results to NASS statistics by county for the years 2002, 2007, 2012, and 2017, shown in Fig. 4. We also compared U-Net predictions of irrigated area by county with FLU estimates of irrigated area by county. These data were available for the years 2009, 2011, 2013, 2015, 2017, 2018, and 2019. Agreement between U-Net, FLU, and NASS was assessed with slope and r^2 from linear regression.

Results and Discussion

The confusion matrix in Table 1 shows the performance of the median model on the validation set. We believe that the number of pixels per class in the validation set more closely represents the actual distribution of irrigated land in Montana compared to data with equal number of pixels per class. Since land cover mapping mostly relies on hand labeled data, the number of examples per class is often chosen to be close to equal. However, the number of pixels in different land classes is usually quite imbalanced; in Montana, for example, forests cover much more of the state by area than irrigated land. If pixels were balanced in the validation set, metrics would represent a skewed picture of how well the model performs. F1-score aggregates the number of true positives, false positives, and false negatives into one metric, defined as $\text{f1}=\frac{TP}{TP+\frac{1}{2}(FP+FN)}$. It's the harmonic mean of precision and recall, and is only close to 1 if both precision and recall are close to 1. We use it to quantify the performance of the U-Net ensemble on the irrigated class, as well as the performance of other maps of irrigation.

		Predicted			
		Irrigated	**Unirrigated**	**Uncultivated**	**Recall**
Actual	**Irrigated**	1788615	184405	104779	0.86
	Unirrigated	114903	104929537	1212801	0.99
	Uncultivated	194506	5334847	105872279	0.95
	Precision	0.85	0.95	0.97	

Table 1: Confusion matrix for median ensemble U-Net results on the validation set. The validation set was not weighted in any way to compensate for class imbalance.

Over all years considered, U-Net had a f1-score for the irrigated class of 0.86 and 0.97 for both the uncultivated and the unirrigated class. There were differing numbers of irrigated labels in different years. The year 2003, for example, only had $\sim 3 \times 10^4$ irrigated pixels and over 2.7×10^7 non-irrigated pixels. The f1-score for 2003 was only 0.4, but for years with less imbalanced distributions (like 2013, with $\sim 3.5 \times 10^5$ irrigated pixels) the f1-score jumped to 0.87. Despite the fact that the f1-score was variable depending on the number of pixels labeled as irrigated per year, the model predictions looked qualitatively similar throughout the time period evaluated.

Value added compared to other products that map irrigation in MT

To the best of our knowledge, IrrMapper, MIrAD-US, and LANID-2012 are the only other high spatial and temporal resolution irrigation maps that cover the state of Montana. MIrAD-US is only available for years 2002, 2007, 2012, and 2017, as it is directly parameterized using NASS statistics, and as such agrees very well with NASS. We did not produce pixel-wise assessments of MIrAD-US maps since they are 250m resolution compared to 30m for LANID, IrrMapper, and U-Net. IrrMapper covers the range 1986-2019, and we were able to access LANID-2008 through LANID-2013. To evaluate the performance of the models (U-Net, IrrMapper-RF, and LANID), we compared each map to the validation split of our ground-truth dataset. To compare bewteen methods, we aggregated, U-Net classes unirrigated and uncultivated into a single class (non-irrigated), as the other products only assign a binary indicator to each pixel.

		OA	P	R	F1
Method	IrrMapper RF	0.993	0.63	0.88	0.73
	LANID RF	0.988	0.49	0.84	0.61
	U-Net	0.997	0.87	0.88	0.88

Table 2: Overall accuracy, precision, recall, and f1 score for the four different maps of irrigation in Montana. Precision, recall, and f1 are only reported for the irrigated class. Statistics were aggregated for the years 2008-2013 for LANID, IrrMapper, and the U-Net ensemble.

The quantitative comparison of the performance of the different mapping techniques is shown in Table 2. Precision and recall are only calculated for the irrigated class. In all metrics considered, the U-Net ensemble outperforms existing maps of irrigation. There is a slight discrepancy between f1 scores in Tables 1 and 2 because labels from years 2003 and 2015 are omitted in Table 2 to keep the comparison between methods based on the same set of pixels, as LANID-US was only available from 2008-2013. The results in Table 2 aren't surprising, as U-Net was trained exclusively on data from Montana. However, it does show that U-Net is capable of classifying irrigated lands with higher precision and recall than other methods in the state of Montana given the training data. Overall accuracy in this case is a misleading metric as the distribution of land cover in Montana is highly imbalanced - all methods have an OA of greater than 98%, despite having different qualitative performance. For other states, especially states where irrigated agriculture is comprises a larger percentage of land cover, overall accuracy may be a better estimate of model performance. F1-score incorporates the errors of omission (precision) and errors of comission (recall) into one metric, and it's clear from Table 2 that U-Net outperforms all other methods in identifying irrigated land. LANID and IrrMapper both have high errors of commission over dryland areas, driving the low values of precision in Table 2.

Validation with NASS and FLU

Fig. 4 shows that U-Net slightly underestimates irrigated area by county when compared to NASS. Fig. 5 indicates that U-Net and FLU data agree fairly well. U-Net estimates more irrigated area compared to FLU data on a county wide level for most of the years compared. Both comparisons show that variance in the comparison between census data (either FLU or NASS) increases with an increase in county irrigated area. The total area of irrigated agriculture predicted by the different methods is shown in Fig. 6. Overall, the median U-Net prediction is much closer to both FLU and NASS than the two other methods examined (LANID and IrrMapper-RF). MIrAD-US agrees very well with NASS, as is expected; the method matches its predictions with NASS statistics.

There is a general increasing trend in irrigated area predictions in all methods of estimated irrigated area. IrrMapper-RF and U-Net agree in the trends in irrigated acreage that they identify; the two lines follow roughly the same shape. IrrMapper-RF predicted irrigated area closely follow trends in precipitation, most likely due to a) confusion in wet areas due to overall higher NDVI and b) the detection of temporary increases in irrigated land because of increased water availability. U-Net does not show the same correlation with precipitation as IrrMapper. In both comparisons to NASS and FLU, r^2 values vary widely, from 0.50 (2002) to 0.73 (2017). The large discrepancy between U-Net and NASS area for certain counties in 2002

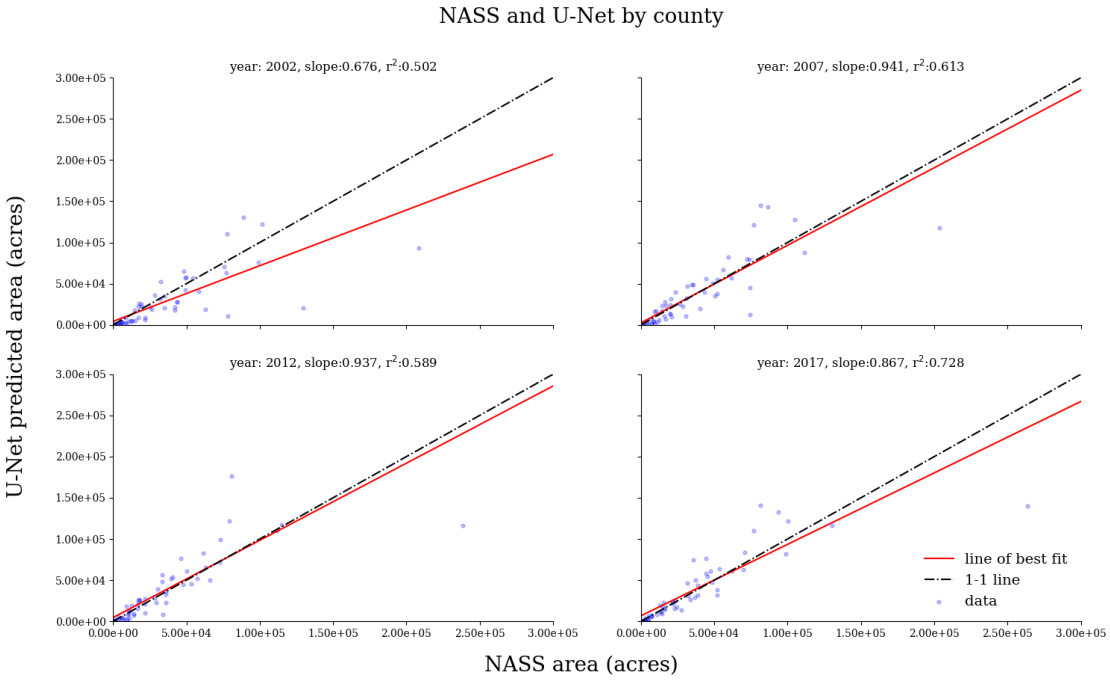

Figure 4: U-Net results compared to NASS statistics. The blue line on the left hand side indicates the line of best fit for the dataset, and the black line the 1-1 line.

is likely attributable to model error in counties on the Rocky Mountain Front. The model erroneously classified areas that it consistently predicts as irrigated in other years as unirrigated. We found that clouds were present in every image save one, meaning the model had little to no information to make classification. The clear outlier in all subplots in Fig. 4 is Beaverhead county, in the bottom right of all plots. NASS reports irrigated area from 200,000 to 260,000 acres, while U-Net reports area from 90,000 to 140,000 acres. FLU data agrees very well with U-Net results for Beaverhead county, with the smallest difference in irrigated area prediction only 2,000 acres. U-Net consistently produces larger estimates of irrigation for a cluster of counties that NASS reports to have around 100,000 acres - Lake, Yellowstone, and Gallatin counties (Fig. 7). FLU estimates of area for the same counties agree closely with U-Net estimates of irrigated area (15,000 acre difference for 2017).

Qualitative performance of models

Inspection of Landsat and NAIP imagery for NASS census years of Beaverhead county indicates that the large discrepancy in irrigated area between NASS and U-Net predictions is most likely due to U-Net omissions of flood irrigated lands as irrigated. U-Net tends to classify flood irrigation or irrigation at the wetlands-agriculture interface as wetlands. It's often hard to tell where wetlands begin and irrigation ends when inspecting high resolution satellite data, so it's not surprising that U-Net has difficulty distinguishing these two classes. In addition, in our training data and our study area generally, the existence of wetlands and irrigation in the same place is possible, and therefore both semantic and physical distinction between irrigated land and wetlands is blurred. This problem may be resolved by more extensive vetting of wetlands training data. Each model has different biases on wetlands. LANID-2012 restricts its classification of irrigated land to land that is not wetlands, impervious surfaces, open water, and forests according to the 2012 CDL. Some flood irrigated land is classified as wetlands in the CDL, resulting in errors of omission. This is evident when inspecting NAIP imagery in the regions that the CDL classifies as wetlands. IrrMapper-RF, on the other hand, tends to label irrigated land that LANID and U-Net miss as irrigated. However, IrrMapper-RF also tends to label true wetlands as irrigated, for example clearly uncultivated land on river islands.

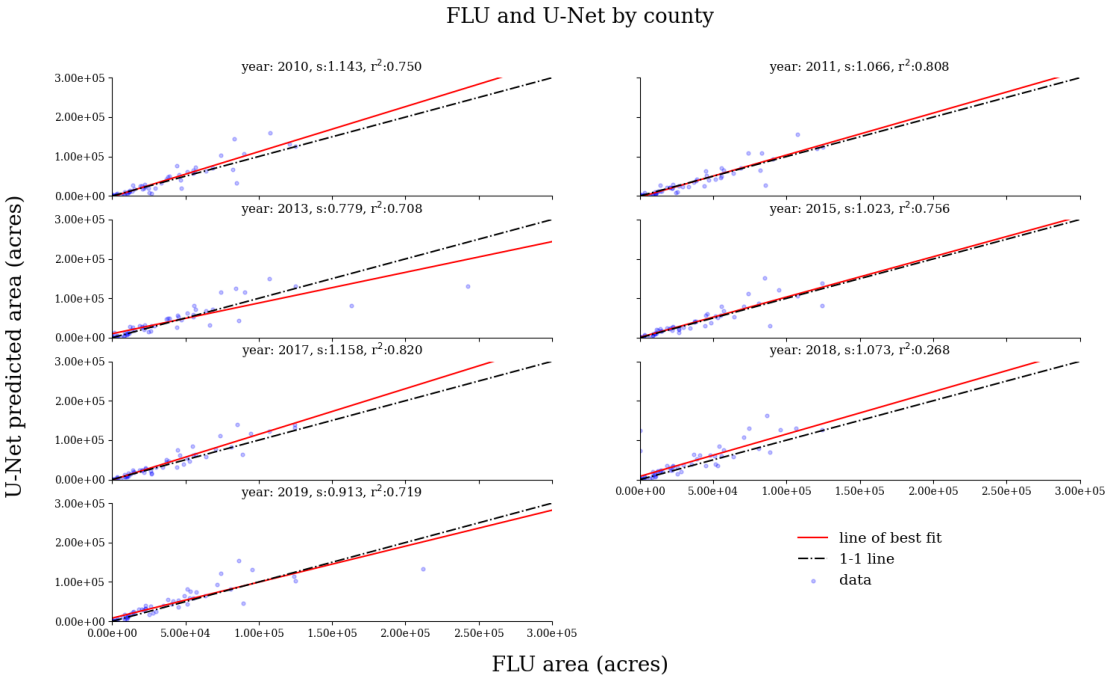

Figure 5: U-Net results vs FLU statistics.

U-Net seems to bridge the gap between the two methods; it labels true wetlands as wetlands and some flood irrigated land as irrigated. However, it does detect as much flood irrigated land as IrrMapper-RF, resulting in lower estimates of irrigated land in counties such as Beaverhead. Some wetlands training data incorporate pixels where flood irrigated land was labeled as wetlands. It's conceivable that the labels were correct at the time of their creation but the land use changed during subsequent years. In general, we expect IrrMapper-RF to overestimate irrigated land near wetlands-agriculture interfaces, U-Net to underestimate somewhat, and LANID to almost entirely ignore. This is evident in the Big Hole Valley near Wisdom, MT where LANID systematically omits irrigated land. Dryland agriculture, in contrast, is very well classified by U-Net compared to other methods. There are few errors of commission. The errors mosty happen at the edges of irrigated fields that are labeled as dryland due to the difference in label and data resolution. LANID has many errors of commission in dryland areas, particularly in the eastern half of Montana. IrrMapper-RF also overpredicts irrigation in dryland agriculture but has fewer errors of commission in dryland areas when compared to LANID.

U-Net systematically overestimates irrigation compared to NASS for Lake, Gallatin, and Yellowstone counties. These are all counties with large areas of irrigated land that reside next to bodies of water, and as such have large irrigation schemes. There are clearly fallowed fields that U-Net is confident are irrigated in some years. Inspection of late summer NDVI reveals that despite U-Net's classification, they are not irrigated during the year inspected. This may be due to the lack of training data describing fallow fields - it was not represented in the training data in the same way as irrigated fields or dryland agriculture. We expect fallow fields to have a more similar spectral signature to irrigated land compared to dryland fields, especially if they're adjacent to irrigated land. In this case, it's not surprising that U-Net results overpredict irrigation in fallowed fields surrounded by irrigation.

Uncertainty in model predictions from prediction IQR

The models in the ensemble did not always predict similar values for each pixel. We formed a per-pixel "uncertainty" by examining the inter-quartile range (IQR) of the softmax predictions of each model. The IQR will be higher when the models disagree on the classification of a given pixel. Ideally, the model

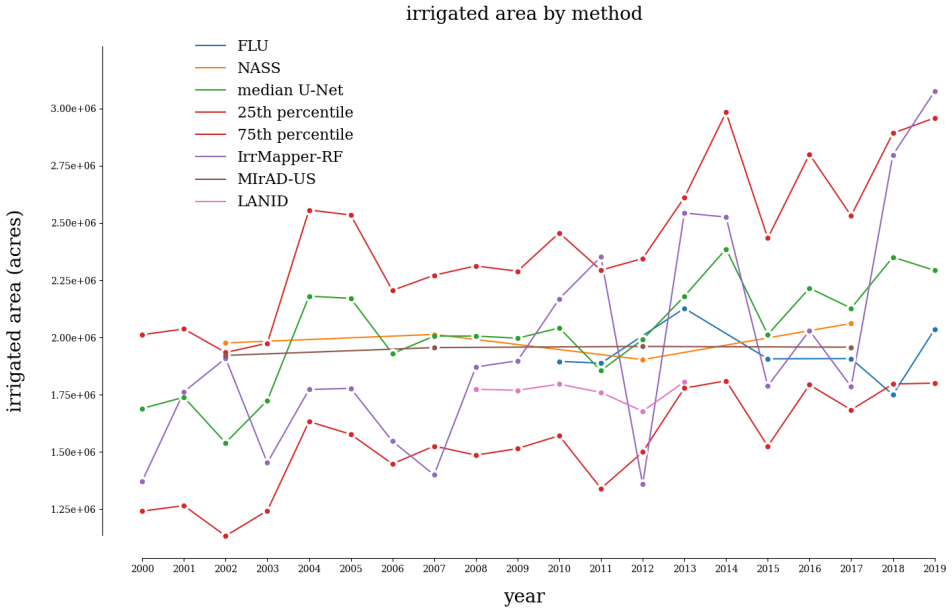

Figure 6: Irrigated area in Montana by method, with NASS and FLU data included. 25th and 75th percentiles for U-Net predictions are shown in red.

ensemble will be less certain when it misclassifies pixels. Histograms of the IQR for misclassified pixels are shown in Fig. 10. The distributions of IQR for errors of commission on the irrigated class (whether on fallowed land, wetlands, dryland agriculture, or forest) are much different than the distributions for errors of commission on other classes. In particular, the pixels that are known to be non-irrigated yet predicted as irrigated tend to have distributions of IQR that peak in the 100-200 range. The 100-200 are quantized softmax probabilities that roughly correspond to a raw probability IQR of 0.4-0.8. The ability of the model ensemble to indicate some uncertainty makes directly fitting NASS or FLU data estimates of area by county possible by considering pixels irrigated that are above or below some IQR threshold. In particular, land near the wetlands-agriculture interface often has a higher IQR than land that is clearly irrigated. Similarly, some fallow land classified as irrigated has a high IQR. The model ensemble also exhibits high IQR in mountainous regions of northwestern Montana that are erroneously classified as irrigated.

The histograms of IQR for all pixels that the models predicted as a certain class are shown in Fig. 11. Despite the fact that there are three orders of magnitude between the number of pixels predicted to be irrigated and non-irrigated, the pixels predicted to be irrigated have a larger median IQR than pixels predicted as other classes. Even removing the first 30 bins (that contain most of the count density for the non-irrigated classes) results in median IQR of 5 and 8 for the unirrigated and uncultivated class, respectively. Median IQR for the irrigated class was 29. Examples of misclassifications that are highly uncertain are shown in Fig. 8. Subfigure a) shows U-Net predicts irrigation in some small regions in the mountains of northwest Montana. The IQR associated with the predictions clearly indicates that the ensemble is uncertain about the irrigated classification. Subfigure b) shows similar results for land that is known to be fallow - the fields outlined in black (which are known to be fallow but labeled by the ensemble as irrigated) clearly have a higher IQR.

The discussion of the qualitative performance of U-Net results relied on maps of irrigation that assigned the median prediction of all 10 models to a given pixel. The 10 models perform differently because they were trained on different subset of the training data. In general, models that performed poorly on the validation set tended to have a less strict definition of irrigation and correctly predicted irrigation at the wetlands-agriculture interface more frequently than the median prediction. In contrast, models that performed well on the validation set tended to have a more restrictive definition of irrigation, and correctly identified more fallowed or dryland fields as irrigated in counties like Lake or Yellowstone. Part of this trend may be

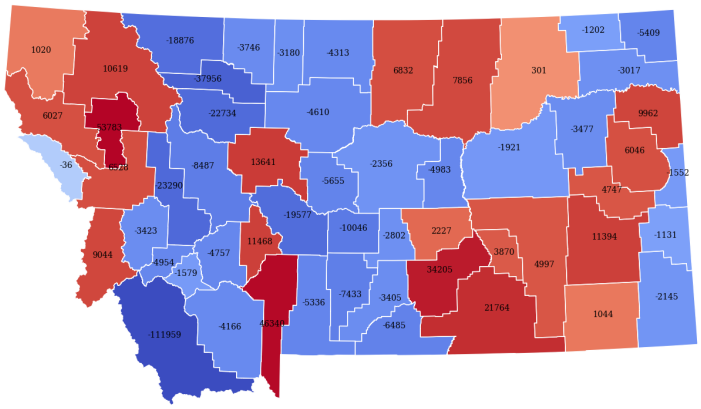

Figure 7: Mean difference in U-Net and NASS area. Mean taken over NASS census years.

explained by the sparsity of labels in the validation set, as models that learn a very restrictive definition of irrigation may not truly reflect reality despite performing well on the validation set. All rasters used in the final analysis are available on GEE, and we recommend examining predictions qualitatively and examining model uncertainty when performing analysis.

Overall, the median predictions include some fallow or unirrigated fields surrounded by irrigated agriculture predicted as irrigated. The final product also tends to miss irrigated land that is irrigated by flood irrigation, such as in Beaverhead county. Despite the qualitative drawbacks of U-Net results, we believe the median predictions are a reasonable representation of irrigation in Montana over the time period studied. In particular, it very rarely labels dryland agriculture as irrigated and is good at distinguishing between true wetlands and irrigation. The different learners in the model ensemble have different characteristics, and it is worthwhile examining the performance of the different models for specific applications. One other advantage of the U-Net ensemble is the performance for years such as 2012 when all data ingested was noisy. In other irrigation mapping products, estimates of irrigated land falsely decrease in 2012 (Fig. 6). U-Net results, on the other hand, don't exhibit the same discontinuity in 2012 as other methods, showing that U-Net learned to gap fill data on its own.

The high accuracy and resolution of U-Net maps allow for a detailed accounting of irrigation in Montana. Overall, U-Net results indicate that irrigated area increased by 603,761 acres from 2000 to 2019. To reduce yearly noise, we averaged the results over 4 non-overlapping 5-year periods, finding that the averaged area over this time increased by 346,141 acres. The year 2002 was excluded from the analysis because of the issues due to missing data. Fig. 9 shows the change in irrigated land between the years 2000 and 2019. Many counties experienced a slight decrease in irrigation, while four strongly increased their irrigated land (Beaverhead, Gallatin, Big Horn, and Madison counties).

Despite the fact that U-Net misclassifies flood irrigated land as wetlands in some areas in Beaverhead county, there is still a visible continuous increasing trend in irrigated agriculture.

Conclusion

We present a new methodology of mapping irrigation (U-Net) at high spatial resolution and over large regions in time and space, applied to the state of Montana. The method used is a deep learning model that ingests temporal mean images of satellite derived spectral data from multiple times throughout the year. It learns to mask clouds and noisy data without explicit guidance. The model is also computationally efficient enough to produce predictions for 20 years over the state of Montana on a personal computer in less than a day. Our

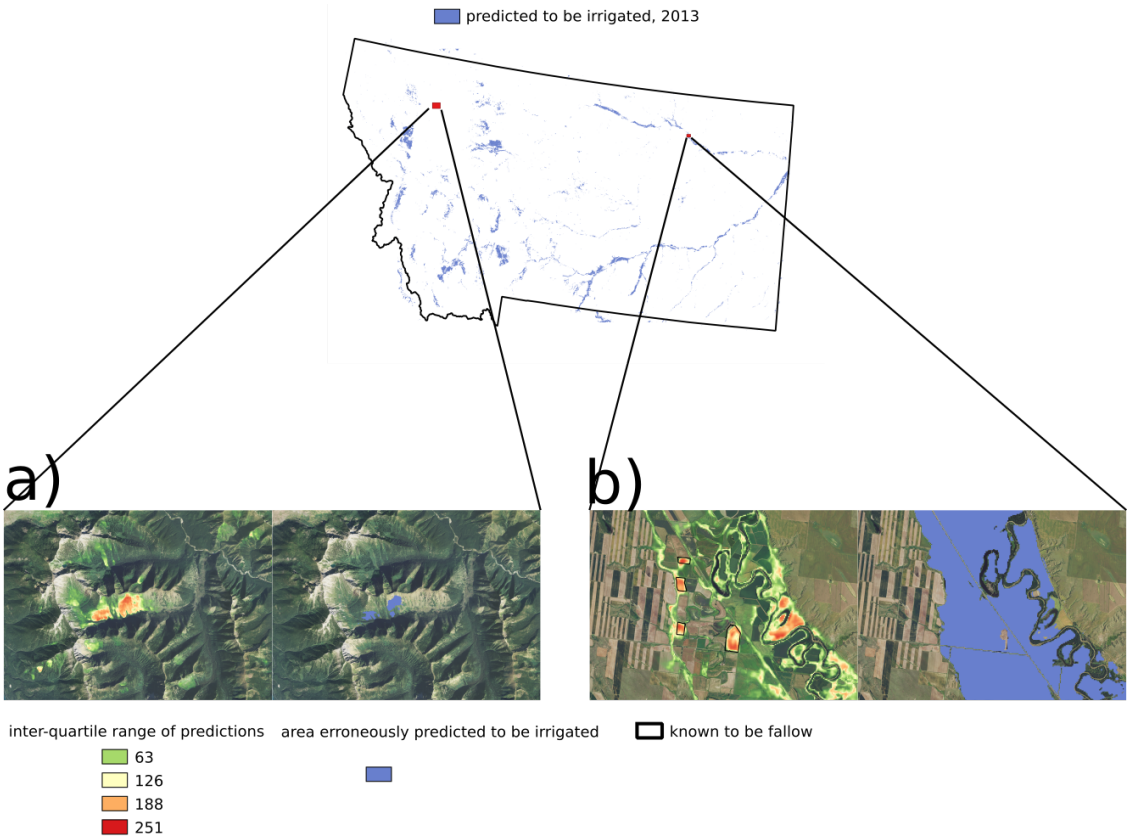

Figure 8: Example of large IQR where the model has incorrectly classified forest and fallowed lands as irrigated.

method agrees better statewide with both NASS and FLU data than existing maps of irrigation in Montana, and outperforms other methods on a hand-labeled validation set. Despite this good performance, U-Net sometimes misclassifies wetlands as irrigated. In addition, the under-representation of fallowed land data in the training set means U-Net predicts irrigation in fallowed fields, especially in regions where irrigation is prevalent. The model results most clearly outperform existing methods in rainfed areas, producing almost no errors of commission in dryland agriculture. The method we propose is the first application of CNNs to mapping irrigation on a large spatial and temporal scale, and was able to be applied with modest compute infrastructure.

Supplementary figures

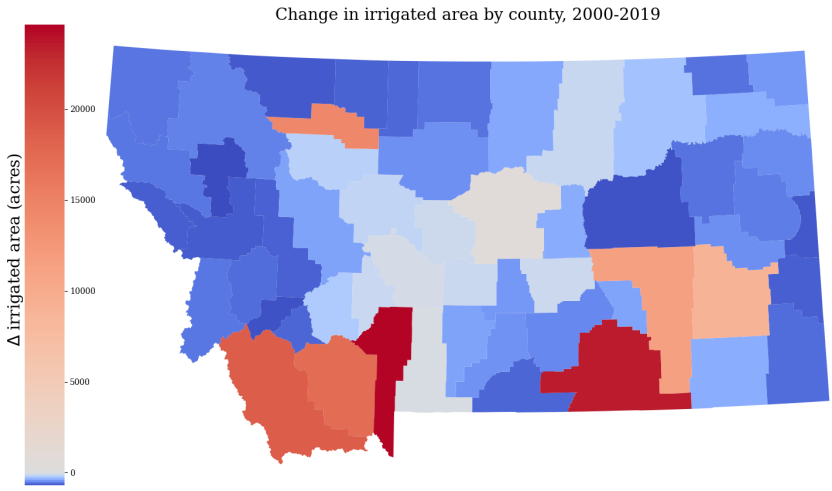

Figure 9: Change in irrigated area by county for the years 2000-2019. To smooth predictions, areas were averaged over 5 year non overlapping periods, and the year 2002 was excluded due to missing data issues.

References

[1] Leo Breiman. "Bagging predictors". In: *Machine Learning* 24.2 (1996), pp. 123–140. ISSN: 08856125. DOI: 10.1023/A:1018054314350.

[2] Jesslyn F. Brown and Md Shahriar Pervez. "Merging remote sensing data and national agricultural statistics to model change in irrigated agriculture". In: *Agricultural Systems* 127 (2014), pp. 28–40. ISSN: 0308521X. DOI: 10.1016/j.agsy.2014.01.004. URL: http://dx.doi.org/10.1016/j.agsy.2014.01.004.

[3] Jillian M. Deines, Anthony D. Kendall, Morgan A. Crowley, et al. "Mapping three decades of annual irrigation across the US High Plains Aquifer using Landsat and Google Earth Engine". In: *Remote Sensing of Environment* 233.October (2019), p. 111400. ISSN: 00344257. DOI: 10.1016/j.rse.2019.111400. URL: https://doi.org/10.1016/j.rse.2019.111400.

[4] Jillian M. Deines, Anthony D. Kendall, and David W. Hyndman. "Annual Irrigation Dynamics in the U.S. Northern High Plains Derived from Landsat Satellite Data". In: *Geophysical Research Letters* 44.18 (2017), pp. 9350–9360. ISSN: 19448007. DOI: 10.1002/2017GL074071.

[5] Kingma Diederik and Jimmy Lei Ba. "ADAM: A Method for Stochastic Optimization". In: *AIP Conference Proceedings* 1631 (2014), pp. 58–62. ISSN: 15517616. DOI: 10.1063/1.4902458. arXiv: 1412.6980v9.

[6] Montana DNRC. *Montana State Water Plan*. Retrieved from http://dnrc.mt.gov/divisions/water/management/docs/state-water-plan/2015_mt_water_plan.pdf. 2015.

[7] FAO. Retrieved from http://www.fao.org/water/en. 2020.

[8] Vivien Sainte Fare Garnot and Loic Landrieu. "Lightweight Temporal Self-Attention for Classifying Satellite Image Time Series". In: (2020), pp. 1–11. arXiv: 2007.00586. URL: http://arxiv.org/abs/2007.00586.

[9] Noel Gorelick et al. "Google Earth Engine: Planetary-scale geospatial analysis for everyone". In: *Remote Sensing of Environment* 202 (2017), pp. 18–27. ISSN: 00344257. DOI: 10.1016/j.rse.2017.06.031. URL: https://doi.org/10.1016/j.rse.2017.06.031.

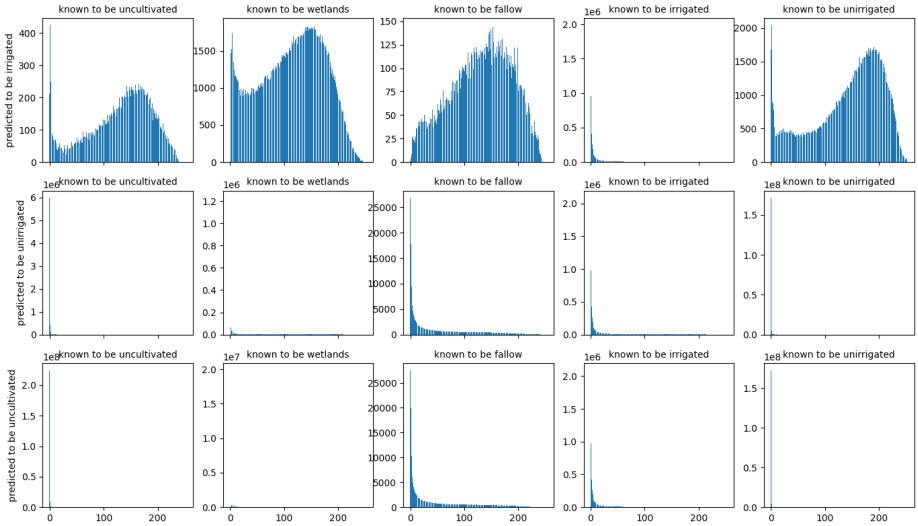

Figure 10: Histograms showing the IQR for correctly and incorrectly classified pixels. Recall the uncultivated and unirrigated classes are aggregated into two separate classes when training the model, composed of forested data and wetlands and dryland agriculture and fallow, respectively. Correctly classified pixels seem to show a much more uniform distribution of IQR compared to misclassified pixels for the pixels that are incorrectly classified as irrigated.

[10] Sergey Ioffe and Christian Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing". In: *Journal of Molecular Structure* 1134 (2015), pp. 63–66. ISSN: 00222860. DOI: 10.1016/j.molstruc.2016.12.061. arXiv: 1502.03167v3.

[11] Milad Janalipour and Mohammad Taleai. "Building change detection after earthquake using multi-criteria decision analysis based on extracted information from high spatial resolution satellite images". In: *International Journal of Remote Sensing* 38.1 (2017), pp. 82–99. ISSN: 13665901. DOI: 10.1080/01431161.2016.1259673. URL: http://dx.doi.org/10.1080/01431161.2016.1259673.

[12] David Ketchum et al. "IrrMapper: A machine learning approach for high resolution mapping of irrigated agriculture across the Western U.S." In: *Remote Sensing* 12.14 (2020), pp. 1–23. ISSN: 20724292. DOI: 10.3390/rs12142328.

[13] Jonas Meier, Florian Zabel, and Wolfram Mauser. "A global approach to estimate irrigated areas - A comparison between different data and statistics". In: *Hydrology and Earth System Sciences* 22.2 (2018), pp. 1119–1133. ISSN: 16077938. DOI: 10.5194/hess-22-1119-2018.

[14] Mutlu Ozdogan and Garik Gutman. "A new methodology to map irrigated areas using multi-temporal MODIS and ancillary data: An application example in the continental US". In: *Remote Sensing of Environment* 112.9 (2008), pp. 3520–3537. ISSN: 00344257. DOI: 10.1016/j.rse.2008.04.010.

[15] Mutlu Ozdogan, Yang Yang, et al. "Remote sensing of irrigated agriculture: Opportunities and challenges". In: *Remote Sensing* 2.9 (2010), pp. 2274–2304. ISSN: 20724292. DOI: 10.3390/rs2092274.

[16] Daifeng Peng, Yongjun Zhang, and Haiyan Guan. "End-to-end change detection for high resolution satellite images using improved UNet++". In: *Remote Sensing* 11.11 (2019). ISSN: 20724292. DOI: 10.3390/rs11111382.

[17] Montana Department of Revenue. *Revenue Final Land Unit (FLU) Classification*. data retrieved from https://mslservices.mt.gov/Geographic_Information/Data/DataList/datalist_Details.aspx?did={18a47176-0d37-406e-981e-570e1b003832}. 2015.

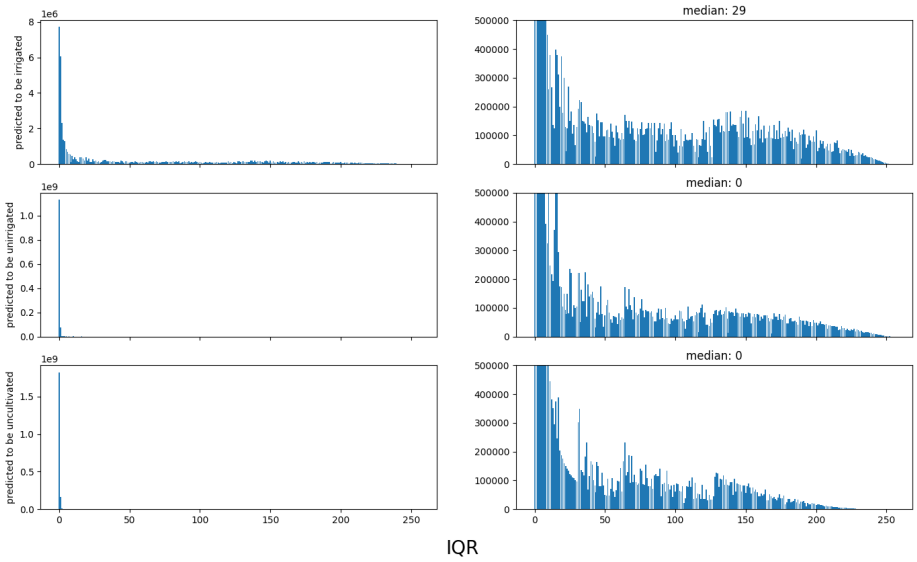

Figure 11: IQR histograms of all pixels predicted to be a given class. The right hand side of the figure shows a view where the y-axis has been restricted. There's a peak in the pixels that are predicted to be irrigated that isn't present to the same degree as in pixels predicted to be other classes.

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 9351 (2015), pp. 234–241. ISSN: 16113349. DOI: 10.1007/978-3-319-24574-4_28. arXiv: 1505.04597.

[19] Marc Rußwurm and Marco K rner. "Multi-temporal land cover classification with sequential recurrent encoders". In: *ISPRS International Journal of Geo-Information* 7.4 (2018), pp. 1–19. ISSN: 22209964. DOI: 10.3390/ijgi7040129. arXiv: 1802.02080.

[20] Marc Rußwurm and Marco Körner. "Self-Attention for Raw Optical Satellite Time Series Classification". In: (2019), pp. 1–20. arXiv: 1910.10536. URL: http://arxiv.org/abs/1910.10536.

[21] J.Meghan Salmon et al. "Global rain-fed, irrigated, and paddy croplands: A new high resolution map derived from remote sensing, crop inventories and climate data". In: *International Journal of Applied Earth Observation and Geoinformation* 38 (2015), pp. 321–334. ISSN: 03032434. DOI: 10.1016/j.jag.2015.01.014. URL: http://dx.doi.org/10.1016/j.jag.2015.01.014.

[22] Marciano Saraiva et al. "Automatic mapping of center pivot irrigation systems from satellite images using deep learning". In: *Remote Sensing* 12.3 (2020), pp. 1–14. ISSN: 20724292. DOI: 10.3390/rs12030558.

[23] Evan Shelhamer, Jonathan Long, and Trevor Darrell. "Fully Convolutional Networks for Semantic Segmentation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.4 (2017), pp. 640–651. ISSN: 01628828. DOI: 10.1109/TPAMI.2016.2572683. arXiv: 1411.4038.

[24] Stefan Siebert et al. "Historical Irrigation Dataset (HID)". In: *Hydrology and Earth System Sciences* 19 (2015), pp. 1521–1545. DOI: 10.13019/M20599. URL: http://dx.doi.org/10.13019/M20599%7B%5C%%7D5Cnhttps://mygeohub.org/publications/8/2.

[25] Prasad S. Thenkabail et al. "Global irrigated area map (GIAM), derived from remote sensing, for the end of the last millennium". In: *International Journal of Remote Sensing* 30.14 (2009), pp. 3679–3733. ISSN: 13665901. DOI: 10.1080/01431160802698919.

[26] National Agricultural Statistics Service USDA. *2017 Census of Agriculture*. Retrieved from https://www.nass.usda.gov/Publications/AgCensus/2017/Full_Report/Volume_1,_Chapter_1_US/usv1.pdf. 2020.

[27] Yanhua Xie et al. "Mapping irrigated cropland extent across the conterminous United States at 30m resolution using a semi-automatic training approach on Google Earth Engine". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 155.July (2019), pp. 136–149. ISSN: 09242716. DOI: 10.1016/j.isprsjprs.2019.07.005. URL: https://doi.org/10.1016/j.isprsjprs.2019.07.005.

[28] Jinfan Xu et al. "DeepCropMapping: A multi-temporal deep learning approach with improved spatial generalizability for dynamic corn and soybean mapping". In: *Remote Sensing of Environment* 247.June (2020), p. 111946. ISSN: 00344257. DOI: 10.1016/j.rse.2020.111946. URL: https://doi.org/10.1016/j.rse.2020.111946.

[29] Chenxiao Zhang et al. "Automatic identification of center pivot irrigation systems from landsat images using convolutional neural networks". In: *Agriculture (Switzerland)* 8.10 (2018). ISSN: 20770472. DOI: 10.3390/agriculture8100147.

[30] Liheng Zhong, Lina Hu, and Hang Zhou. "Deep learning based multi-temporal crop classification". In: *Remote Sensing of Environment* 221.November 2018 (2019), pp. 430–443. ISSN: 00344257. DOI: 10.1016/j.rse.2018.11.032. URL: https://doi.org/10.1016/j.rse.2018.11.032.